

# An improved EM algorithm for remote sensing classification

YANG HongLei<sup>1</sup>, PENG JunHuan<sup>1\*</sup>, XIA BaiRu<sup>2</sup> & ZHANG DingXuan<sup>1</sup><sup>1</sup> School of Land Science and Technology, China University of Geosciences, Beijing 100083, China;<sup>2</sup> School of Engineering and Technology, China University of Geosciences, Beijing 100083, China

Received May 16, 2012; accepted August 30, 2012; published online October 20, 2012

The use of a general EM (expectation-maximization) algorithm in multi-spectral image classification is known to cause two problems: singularity of the variance-covariance matrix and sensitivity of randomly selected initial values. The former causes computation failure; the latter produces unstable classification results. This paper proposes a modified approach to resolve these defects. First, a modification is proposed to determine reliable parameters for the EM algorithm based on a *k*-means algorithm with initial centers obtained from the density function of the first principal component, which avoids the selection of initial centers at random. A second modification uses the principal component transformation of the image to obtain a set of uncorrelated data. The number of principal components as the input of the EM algorithm is determined by the principal contribution rate. In this way, the modification can not only remove singularity but also weaken noise. Experimental results obtained from two sets of remote sensing images acquired by two different sensors confirm the validity of the proposed approach.

**Gaussian mixture model, EM algorithm, Kernel density estimation, principal component transformation**

**Citation:** Yang H L, Peng J H, Xia B R, et al. An improved EM algorithm for remote sensing classification. *Chin Sci Bull*, 2013, 58: 1060–1071, doi: 10.1007/s11434-012-5485-4

The EM algorithm, first proposed by Dempster et al. [1], is an approach to the iterative computation of maximum likelihood (ML) estimates or maximum a posterior (MAP) estimates and has been broadly used for tasks in remote sensing image analysis, such as remote sensing classification [2–8], segmentation [9], image fusion [10], change detection [11–13] and quantitative remote sensing [14].

The Gaussian mixture model is a weighted sum of Gaussian probability density functions (referred to as Gaussian components) of the mixture model describing a class, and it has been widely used in pattern recognition and classification [15–18]. In remote sensing classification, a unimodal assumption for class conditional distribution is unsuitable for remote sensing images, particularly for high spatial resolution images. It is appropriate to describe the class conditional distribution as a Gaussian mixture model. Without a model or a classification label, it would be difficult to determine the parameters of a Gaussian mixture

model in classification applications. Newton-Raphson and scoring algorithms in parameter solutions are very complex and difficult [19,20]. The EM algorithm-constructed maximum likelihood has good properties because of effective model labels [1,21]. However, the EM algorithm is very sensitive to initializations and easily gets trapped in local minima. In practice, the algorithm runs many times with different initial parameters, and various local search heuristics are used to find better parameters near convergences. Moreover, significant difficulty is encountered in estimating parameters of the Gaussian mixture models with creasing dimensions. Multi-spectral and hyper-spectral images with high correlations can lead to a singular variance-covariance matrix that terminates in an iteration of the EM algorithm without a reliable result [8,22,23]. Furthermore, in the case of a large component overlap, the EM algorithm suffers from slow convergence [24]. Finally, the EM algorithm fails when the covariance matrix corresponding to one or more components becomes ill-conditioned (singular or nearly singular).

\*Corresponding author (email: pengjunhuan@163.com)

Recently, many researchers have proposed several modified techniques for the EM algorithm in remote sensing applications. Fwu and Djuric [25] described an initialization procedure based on a tree-structured scheme that provided excellent initial estimation. Fraley and Raftery [26] proposed the use of the Bayesian information criterion (BIC) as the criterion statistic in EM clustering. The  $k$ -means algorithm [27] can also provide initial values for the EM algorithm. The class number is not available, and the random initial values are unknown, leading to larger differences in classification results. Figueiredo and Jain [3] proposed a finite mixture model based on an unsupervised algorithm by combining estimation and selection into a single algorithm. Luo et al. [4] stated that the improved EM algorithm manually determined the initial values for the EM algorithm. An adaptive EM algorithm, taken from Chakravarty et al. [5], can adapt to capture local statistics and fit in the non-stationary case. Bazi et al. [28] proposed an image threshold method in EM algorithms that can use a robust initialization based on genetic algorithms. Tyagi et al. [29] presented a multistage clustering technique for unsupervised classification based on the graph-cut initialization of the EM algorithm. Wen et al. [30] described a novel method to improve the global optimality and classification performance consistency of synthetic aperture radar (SAR) imagery by integrating a genetic algorithm (GA) with an EM algorithm. Ari and Aksoy [31] presented a novel clustering algorithm based on particle swarm optimization (PSO) for the maximum likelihood estimation of Gaussian mixture models to prevent the effective use of population-based algorithms in clustering problems. As shown by Korting and Dutra [7], a modified version of the standard EM clustering algorithm can prevent clustering when centers are too close and can improve EM initialization by providing results of the well-known  $k$ -means algorithm as seed points. However, as for the sensitivity of the  $k$ -means algorithm to initializations, the classification results are always unreliable. Although these methods use different strategies to initialize EM algorithms to avoid being trapped in local minima, they are not suitable for multi-spectral and hyper-spectral images with high spectral correlations. Moreover, they do not address the singular covariance matrix in iterations.

In an effort to solve these problems with the EM algorithm, a new approach that takes the logarithm of multi-spectral or hyper-spectral images to highlight or reinforce the differences between classes was developed. The principal component transformation is applied to the log-transformed data to obtain a set of uncorrelated data. Then, kernel density estimation can be used to determine the probability density function of the first principal component. According to the peak of the probability density function, the class number and initial centers are determined for the  $k$ -means algorithm. The classification results are obtained using the  $k$ -means algorithm based on the first principal component. Initializations of the multi-dimensional EM

algorithm can be efficiently determined using the classification results. The EM algorithm is used to perform an iterative procedure to determine the parameters of the Gaussian mixture model. Finally, the classification map is obtained by a Bayesian classifier. In addition, an ill-condition covariance matrix, which affects the EM algorithm in iteration, can be detected and changed into a valid covariance matrix.

## 1 The improved EM algorithm

### 1.1 Gaussian mixture model and the EM algorithm

Suppose that we have  $p$ -dimensional vector data  $x$  ( $x \in R^p$ ) from a mixture of  $g$  component Gaussian distributions. Then, the density function is defined as [19]:

$$f(x; \theta) = \sum_{i=1}^g \alpha_i f_i(x; \theta_i), \quad (1)$$

where  $\theta = \{\theta_1, \dots, \theta_g\}$  are the unknown parameters for each  $f_i(x; \theta_i)$ ,  $i = 1, \dots, g$ , which are Gaussian density functions. Additionally,  $g$  is assumed to be the class number. Here,  $\alpha_i \in [0, 1]$  are the mixing proportions subject to:

$$\sum_{i=1}^g \alpha_i = 1, (\alpha_i \geq 0, i = 1, \dots, g). \quad (2)$$

The parameters  $\theta_i = \{\mu_i, V_i\}$  in such a model are the means,  $\mu_i$ , and the covariance matrices,  $V_i$ . The entire set of parameters for the Gaussian mixture model is  $\theta = (\alpha_1, \dots, \alpha_g; \theta_1, \dots, \theta_g)$ .

By estimating the parameters of the Gaussian mixture model to achieve the purpose of classification, it is proven that this method is very effective. For an observational data set  $\{x_1, x_2, \dots, x_n\}$ , the maximum likelihood estimation (MLE) of  $\theta$  can be determined by

$$\theta_{MLE} = \arg \max_{\theta} \prod_{i=1}^n f(x_i; \theta). \quad (3)$$

Eq. (3) is an incomplete-data problem lacking observational information [32]. Due to a lack of knowledge about the data, it is very complicated to solve for  $x_i$  directly (eq. (3)).

The EM algorithm is a widely used approach to the iterative computation of MLEs applied to a variety of incomplete data problems. The principles are as follows [33–35]:

Let  $z_i = (z_{ij})$ ,  $j = 1, 2, \dots, g$ , be the label vector of the observational data set:

$$z_{ij} = \begin{cases} 1 & x_i \in \omega_j, \\ 0 & x_i \notin \omega_j. \end{cases} \quad (4)$$

The observational data set and the label vector together provide a so-called complete data set,  $y = (X; z) = (x_1,$

$x_2, \dots, x_n; z_1, z_2, \dots, z_n$ ), and the following likelihood function,

$$L_c(y, \theta) = \prod_{i=1}^n \prod_{j=1}^g (P(x_i \in \omega_j) f_i(x_i / \omega_j))^{z_{ij}} \\ = \prod_{i=1}^n \prod_{j=1}^g (\alpha_j f_i(x_i / \theta_j))^{z_{ij}}, \quad (5)$$

corresponds to the following log likelihood function,

$$\log L_c(y, \theta) = \sum_{i=1}^n \sum_{j=1}^g z_{ij} (\log \alpha_j + \log f_i(x_i / \theta_j)). \quad (6)$$

The EM algorithm consists of two steps: the E step, or the expectation step, and the M step, or the maximization step. In the E step, let  $\theta^{(0)}$  be the initial value for  $\theta$ , and then begin the first iteration, which will compute the conditional expectation of the complete data, which is

$$S(\theta, \theta^{(k)}) = E(\log L_c(y; \theta) / X, \theta^{(k)}) \\ = \sum_{i=1}^n \sum_{j=1}^g E(z_{ij} / x_i, \theta^{(k)}) (\log \alpha_j + \log f_i(x_i / \theta_j)). \quad (7)$$

The M step finds the  $(k+1)$ th estimation  $\theta^{(k+1)}$  of  $\theta$  by maximizing  $S(\theta, \theta^{(k)})$ , where

$$z_{ij}^{(k)} = E(z_{ij} / x_i, \theta^{(k)}) = \sum_{j=1}^g z_{ij} P(\omega_j / x_i, \theta^{(k)}) \\ = P(\omega_j / x_i, \theta^{(k)}) \\ = \frac{P(x_i \in \omega_j / \theta^{(k)}) f_j(x_i / \omega_j, \theta^{(k)})}{\sum_{j=1}^g P(x_i \in \omega_j / \theta^{(k)}) f_j(x_i / \omega_j, \theta^{(k)})}, \quad (8)$$

$$\alpha_i^{(k+1)} = n^{-1} \sum_{j=1}^g z_{ij}^{(k)}, \quad (9)$$

$$\mu_i^{(k+1)} = \frac{\sum_{j=1}^g z_{ij}^{(k+1)} x_j}{\sum_{j=1}^g z_{ij}^{(k+1)}}, \quad (10)$$

$$V_i^{(k+1)} = \frac{\sum_{j=1}^g z_{ij}^{(k+1)} (x_j - \mu_i^{(k+1)}) (x_j - \mu_i^{(k+1)})'}{\sum_{j=1}^g z_{ij}^{(k+1)}}. \quad (11)$$

Repeat the E step and the M step until the criterion  $\|\theta^{(k+1)} - \theta^{(k)}\| \leq \varepsilon$  is met, where  $\varepsilon$  is an arbitrary, small positive threshold.

## 1.2 Kernel density estimation

In statistics, kernel density is a non-parametric method for

estimating the probability density function of a random variable. If  $x_1, \dots, x_n \sim f$  is an independent and identically distributed sample of a random variable, then the kernel density approximation of its probability density function is [36]:

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right), \quad (12)$$

where  $K$  is a kernel and  $h$  is a smoothing parameter. Typically,  $K$  is taken to be a standard Gaussian function with a mean of zero and a variance of 1. Thus, the variance is controlled indirectly through the parameter  $h$ . Many of the results discussed in the literature concern bandwidth determination for kernel estimation. The bandwidth,  $h$ , applied in this paper is close to the optimal value for a Gaussian density, which follows that of Bowman and Azzalini [36],

$$h = \delta \left( \frac{4}{3n} \right)^{\frac{1}{5}}, \quad (13)$$

where  $n$  is the sample number and  $\delta$  is the sample standard deviation.

## 1.3 Principal component transformation

Multi-spectral data or hyper-spectral images in which the spectra are highly correlated will lead to a singular variance-covariance matrix and EM algorithm termination. Large amounts of data will reduce the efficiency of the classification algorithm. The application of the principal component transformation [37,38] to correlated remote sensor data will result in an uncorrelated data set and will effectively separate redundancy from the data. Accordingly, this procedure can guarantee the stability of the classification algorithm and can improve the classification efficiency. The procedure is

(1) For the data set  $X = (x_1, x_2, \dots, x_n)$ , compute the

mean  $\bar{\mu} = n^{-1} \sum_{i=1}^n x_i$  and the residual  $v_i = x_i - \bar{\mu}$ .

(2) Calculate the covariance matrix:  $\text{cov}(x) = (n-1)^{-1} \sum_{i=1}^n v_i v_i^T$ .

(3) Use singular value decomposition on the covariance matrix,  $\text{cov}(x) = USU^T$ , where  $U$  is the orthogonal matrix and  $S$  is the diagonal matrix whose diagonal elements are eigenvalues.

(4) Compute the principal component analysis:  $Z_i = U^T x_i$ .

The input for the EM classification algorithm is the first  $k$  principal component, not the raw multi-spectral image, which can compress the raw data and eliminate the singularity of the variance-covariance matrix. In addition, logarithmizing the original image prior to the principal component transformation can enhance the class differences,

which can improve classification accuracy.

#### 1.4 The improved EM classification algorithm

The proposed method is structured in three phases intended to: (1) Initialize the parameters of the Gaussian mixture model for the EM algorithm using the  $k$ -means clustering algorithm based on the first principal component; (2) use the EM algorithm to perform the iterative procedure to determine parameters of the Gaussian mixture model; and (3) classify each pixel in the image with a Bayesian classifier. In the following section, each phase is described in detail. A block scheme describing the proposed algorithm is shown in Figure 1.

Phase 1: To provide reliable initial parameters characterizing the Gaussian mixture model, a logarithm is applied to multi-spectral or hyper-spectral images. The principal component transformation described in section 1.3 is applied to the log-transformed data, and the kernel density estimation is employed for the principal component. The class number and initial centers are identified according to the peak of the probability density function. Finally, the  $k$ -means clustering algorithm is used to partition pixels of the image based on the principal component. The initial centers  $\mu$  and variance-covariance matrix  $V$ , as well as the

set of weights  $\alpha$  for the EM algorithm are determined on the basis of the number of samples contained in the corresponding cluster.

Phase 2: This module consists of the iterative procedure for probability estimation. The EM algorithm is employed over all of the image pixels. The EM algorithm allows the maximum likelihood (ML) estimator of the parameters to characterize a certain distribution in the presence of incomplete observations.

First, the initial parameters for EM obtained from the  $k$ -means algorithm are applied to the EM approach. Thus, the number of iterations and the computational time are reduced. Then, in accordance with eqs. (8)–(11), the updated estimates for the unknown parameters are obtained. With each iteration, the set of estimated parameters provides an increased log-likelihood until a local maximum is reached. Convergence is reached when the relative increase in the log-likelihood is lower than a prefixed threshold  $\varepsilon$ , i.e.

$$\|\theta^{(k+1)} - \theta^{(k)}\| \leq \varepsilon. \quad (14)$$

The EM algorithm tends to converge towards a singular solution, especially when the data are (nearly) insufficient or correlative. With every iteration, the singular of the covariance matrix  $V$  is tested and forced to be a valid matrix.

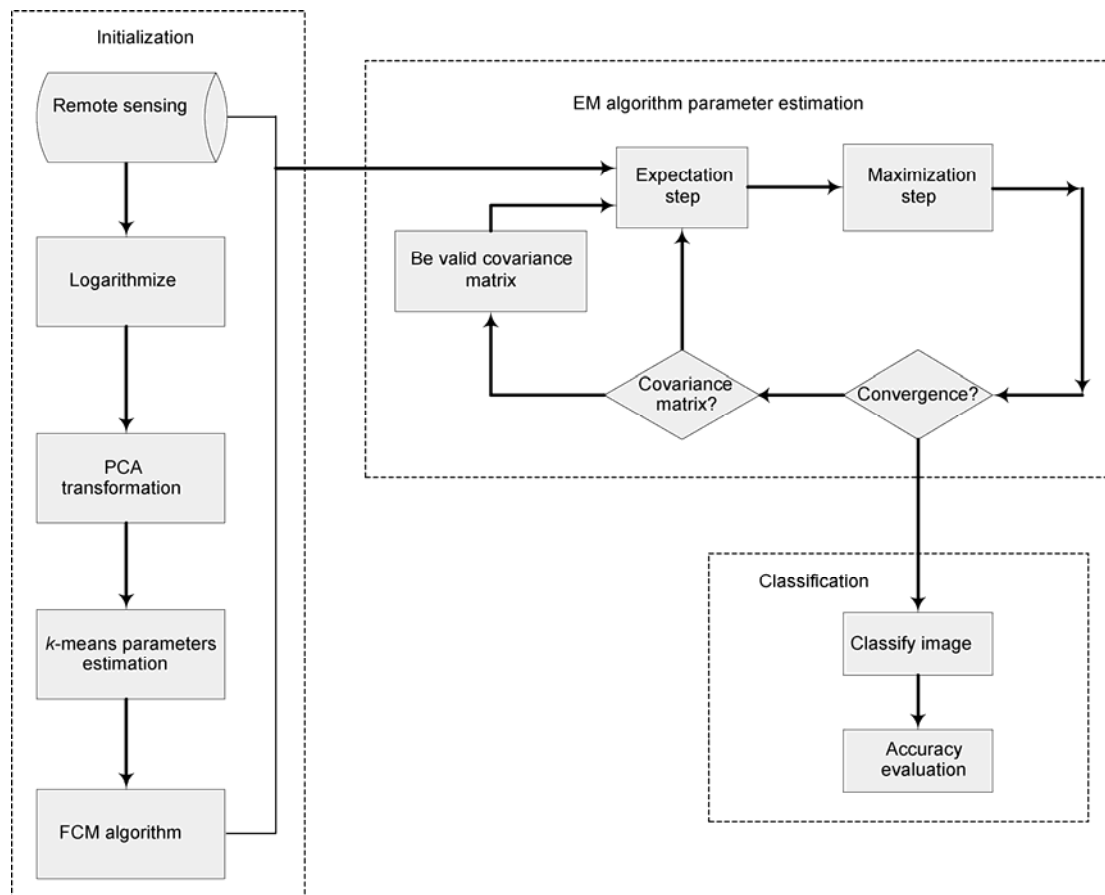


Figure 1 Block scheme describing the proposed technique.

If the input matrix is already valid, it is returned unchanged. When the covariance matrix is ill-conditioned, it is forced into a complex conjugate symmetric as follows:

$$V \leftarrow V - \frac{V - V^*}{2}, \quad (15)$$

where  $V^*$  is the complex conjugate matrix. Possible imaginary components are removed from the diagonal, which results in a Hermitian matrix. This matrix is tested for positive definiteness. If the matrix fails the test, and there is a negative element in the diagonal, then one percent of the maximum element in the diagonal is added to all of the diagonal elements. If the matrix fails the test and there is not a negative element in the diagonal, then all of the diagonal elements are increased by one percent.

By growing the covariance matrix diagonal, the singular matrix can be avoided.

Phase 3: The estimates obtained at convergence can be used to obtain the final classification map based on the Bayes classifier [39]. Once the final classification map is obtained, the classification accuracy can be evaluated based on the Kappa coefficient [40–42].

## 2 Results and analysis

Several experimental trials were conducted to assess the performance of the proposed approach. For these trials, two data sets corresponding to a QuickBird image of Maricopa County, Arizona, USA, and an ASTER image of Beijing, China, were considered.

There were two main objectives of our experimental analysis: (1) To test the robustness of the proposed approach for choosing class number and initial centers; and (2) to demonstrate that the accuracies obtained by the proposed method are comparable to the  $k$ -means algorithm and the modified EM algorithm. In the following section, the promising results for each of the two study objectives are discussed in detail.

### 2.1 Maricopa study area

The first data set used in the experiment was acquired by the QuickBird satellite. The image (250 pixel  $\times$  250 pixel) covers Maricopa County, Arizona, USA, on March 17, 2004 and includes a multi-spectral image in 4 bands with a spatial resolution of 2.4 m and a panchromatic image with a resolution of 0.6 m. The area mainly consists of a green belt, water, roads, residential buildings and bare land, as shown in

Figure 2.

In this experiment, the panchromatic image was resampled to a spatial resolution of 2.4 m so that it could be combined with the multi-spectral image. The skewness of the raw data and the log-transformed data were calculated and are listed in Table 1. As observed in Table 1, the normality of the raw data was corrected by logarithmizing. Figure 3 describes the behavior of the  $k$ -means algorithm with different initial centers. The  $k$ -means algorithm is sensitive to the initial centers. When addressing random initial centers close to the final solution, the  $k$ -means algorithm has high likelihood of finding the cluster centers. Otherwise, it will lead to incorrect clustering results. As a result, initializing the EM algorithm using  $k$ -means solutions is unreliable. In this paper, the class number was fixed at 6, and the corresponding initial centers were set to  $((-0.77, -0.27); (0.08, 1.09); (0.14, 0.81); (1.12, -0.24); (2.16, -0.55); (2.52, -0.6))$  based on the density function of the first two principal components (Figure 4). No prior knowledge was assumed to be available for the initialization of the proposed technique, so the performance of the proposed method was assessed under rather critical conditions. The result obtained by the EM algorithm is shown in Figure 5. The obtained estimates of the density functions for the classes accurately describe the behavior of the histogram of the first principal component. Therefore, choosing  $k=6$  is reasonable based on Figure 6. The Bayesian information criterion (BIC) was also used to choose the number of components.  $k=6$  is obtained from Figure 7, which shows the BIC plot for the first

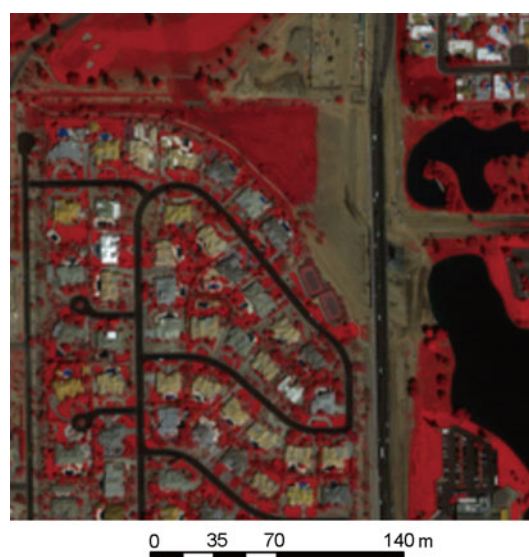
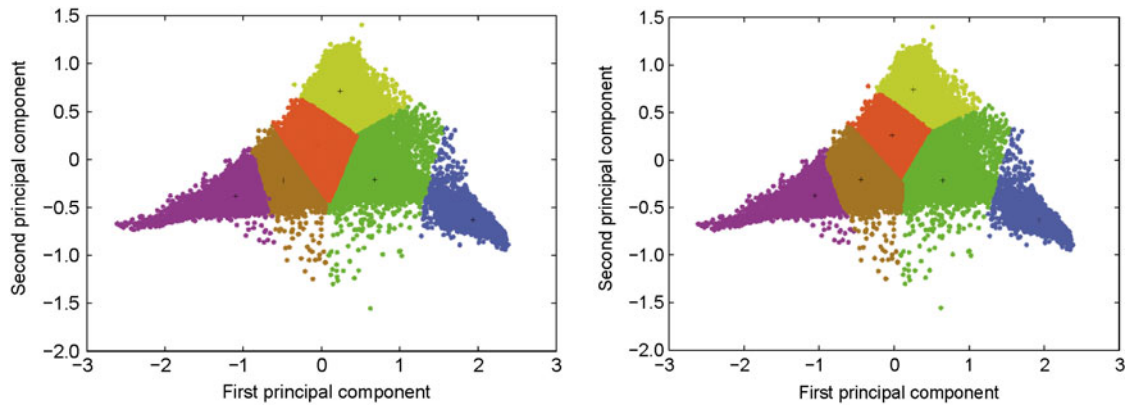


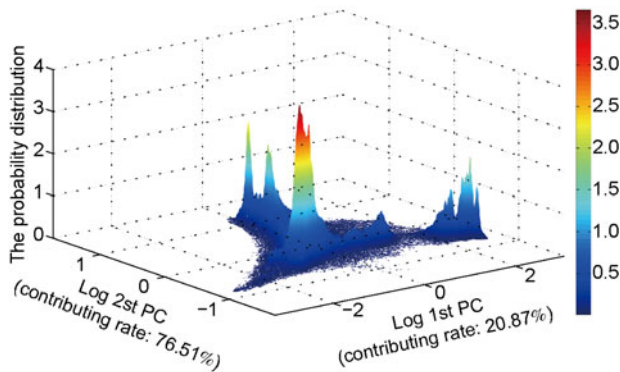
Figure 2 Pseudo color composite of bands 4, 3, and 2.

Table 1 Skewness of raw data and logarithmized data

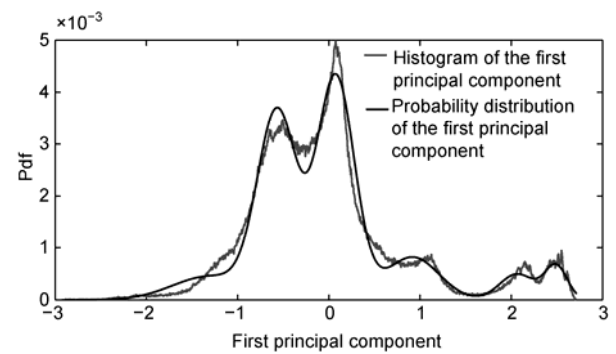
	Panchromatic image	Band 1	Band 2	Band 3	Band 4
Skewness of raw data	-0.97	1.97	1.54	0.88	-1.83
Skewness of logarithmized data	0.59	0.67	0.32	-0.08	-0.43



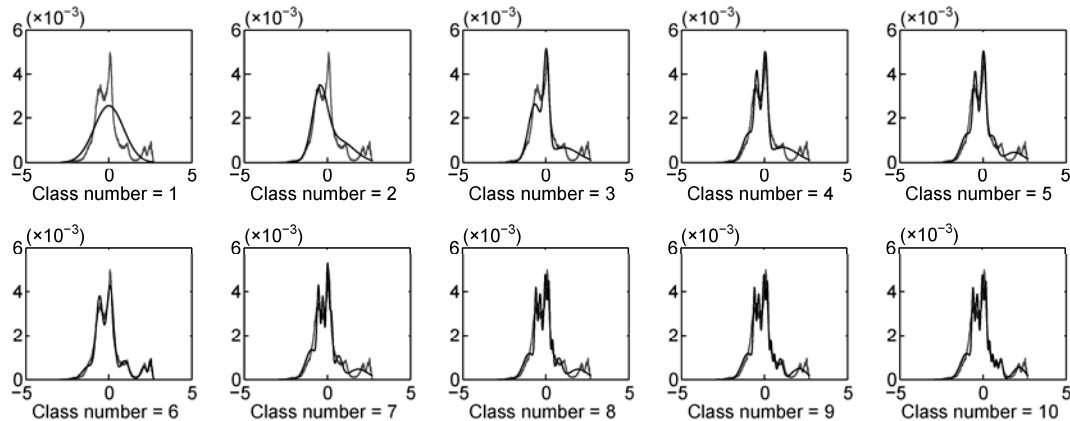
**Figure 3** *k*-means results using different centers.



**Figure 4** Density function of the first two principal components.



**Figure 5** Comparison of the histogram to the probability distribution of the first principal component. The probability distribution was obtained by the technique proposed in this manuscript. The histogram was normalized to permit a direct comparison.



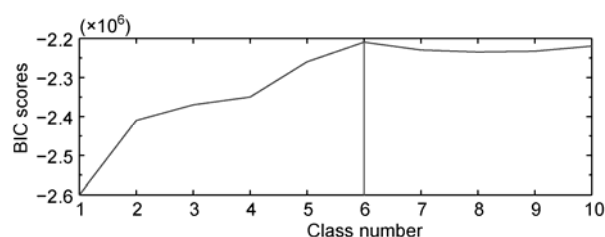
**Figure 6** Histogram of the probability distribution of the first principal component. The probability distribution as a function of the class number  $k = 1, \dots, 10$ . The histogram was normalized to permit a direct comparison. The  $x$ -axis is the principal component, and the  $y$ -axis is the probability density function.

principal component and that it is consistent with the proposed technique.

The classification map (six classes: road, water, building, bare land 1, bare land 2 and green belt) provided by the proposed technique was compared with the classification map obtained by applying the *k*-means and modified EM

algorithms. The first and second principal components, which include 97% of the raw data (shown in Figure 4) as input data, were used in the trial. The same initial values were set for the *k*-means algorithm and the proposed algorithm. The classification results from the *k*-means algorithm and the proposed algorithm used the same initial values and



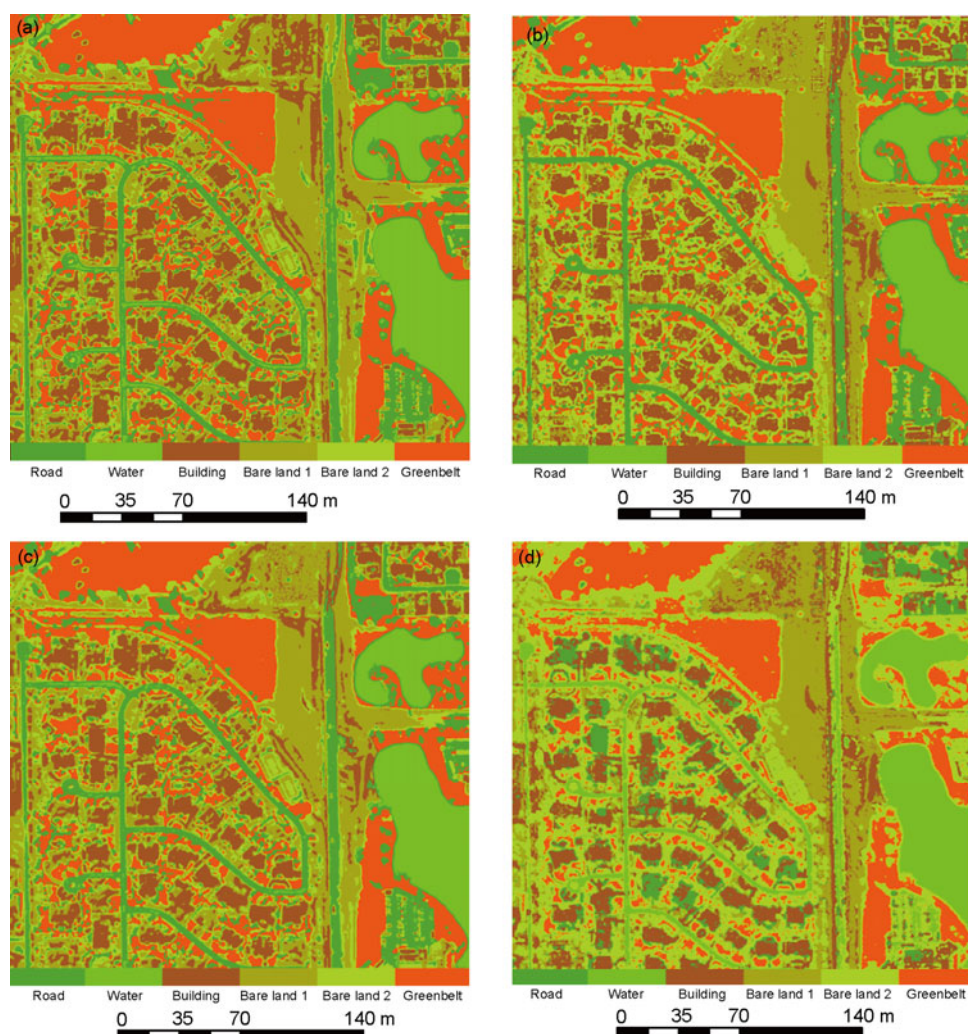


**Figure 7** A plot of BIC scores as a function of the class number for the first principal component.

are shown in Figure 8(a) and (b). The best classification result is shown in Figure 8(c); this result was obtained from 60 tests of the modified EM algorithm. The proposed and modified EM algorithms yield better results for road and bare land 1 (cement pavement) by observing these figures. The results of the proposed and the modified algorithm were similar. Five hundred ground truth points were randomly sampled on the reference image (Figure 2) without any consideration of informational class distribution to avoid statistical bias. These points were classified by visual

interpretation. The overall accuracy and Kappa coefficient were calculated and are listed in Tables 2–4. As observed in Tables 2–4, the proposed algorithm and the modified EM algorithm both produced a higher classification accuracy than the *k*-means algorithm. The result of the proposed algorithm and the modified EM algorithm were basically consistent. In this experiment, the classification map obtained by using the raw data is shown in Figure 8(d). By comparing Figure 8(b) and (c), road and water cannot be classified due to the similarity between the spectra. Before a principal component transformation was applied, a logarithmic transformation was used on the raw data to enhance the low-intensity pixel values and to compress the high-intensity values into a relatively small pixel range, which improved the normality (Table 1) and made the transformed data satisfy the maximum likelihood classification based on a Gaussian distribution. As a result, we were able to successfully classify road and water using our methodology.

The density function of the third and fourth principal components were unimodal noises. Thus, the density function of these components can result in classification errors.



**Figure 8** (a) The *k*-means result; (b) the proposed algorithm result; (c) the modified EM result; (d) the proposed algorithm result.

**Table 2** Classification accuracy of the proposed algorithm<sup>a)</sup>

	Water	Road	Building	Bare land 1	Bare land 2	Green belt	Sum
Water	82	0	0	0	0	0	82
Road	0	42	0	2	4	1	49
Building	0	0	68	2	3	1	74
Bare land 1	0	3	5	85	12	1	106
Bare land 2	0	3	4	14	73	3	97
Green belt	0	2	3	7	11	69	92
Sum	82	50	80	110	103	75	500

a) Overall classification accuracy: 83.8%; Kappa coefficient: 80.37%.

**Table 3** Classification accuracy of the modified EM algorithm<sup>a)</sup>

	Water	Road	Building	Bare land 1	Bare land 2	Green belt	Sum
Water	82	0	0	0	0	0	82
Road	0	41	1	4	2	1	49
Building	0	2	62	4	5	1	74
Bare land 1	0	2	7	86	10	1	106
Bare land 2	0	7	4	15	69	2	97
Green belt	0	0	7	6	13	66	92
Sum	82	52	81	115	99	71	500

a) Overall classification accuracy: 81.2%; Kappa coefficient: 77.23%.

**Table 4** Classification accuracy of the *k*-means algorithm<sup>a)</sup>

	Water	Road	Building	Bare land 1	Bare land 2	Green belt	Sum
Water	82	0	0	0	0	0	82
Road	0	40	2	4	2	1	49
Building	0	2	60	6	5	1	74
Bare land 1	0	4	8	78	15	1	106
Bare land 2	0	8	7	20	59	3	97
Green belt	0	0	7	8	17	60	92
Sum	82	54	84	116	98	66	500

a) Overall classification accuracy: 75.8%; Kappa coefficient: 70.71%.

However, the first and second principal component as input data can remove noise and increase the reparability.

## 2.2 Beijing study area

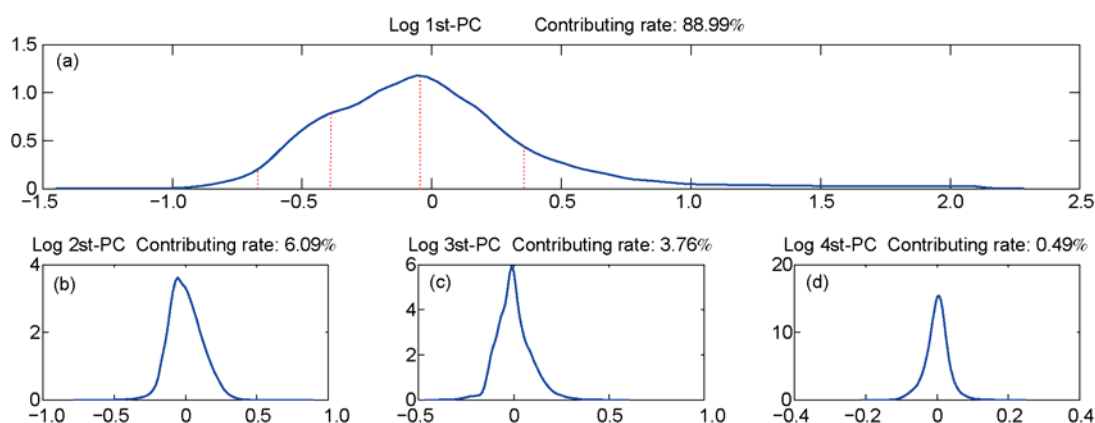
The second data set used in the experiment was collected by ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer Instrument Aboard Terra) on April 9, 2004 with 15 bands from three telescopes. Of the 15 bands, 4 were in the visible and near infrared (VNIR) spectrum with a 15 m spatial resolution, 6 were in the short wave infrared (SWIR) spectrum with a 30 m spatial resolution, and 5 bands were in the thermal infrared (TIR) spectrum with a 90 m spatial resolution. Moreover, ASTER is a Level-1B product from the Land Processes DAAC company, which implies that the radiometric calibration and geometric correction coefficients were applied. In this experiment, only the VNIR and SWIR image bands were used, as the TIR data have a much lower spatial resolution and cannot provide useful information for classification. The SWIR images

were resampled to a spatial resolution of 15 m to match the VNIR image special resolution.

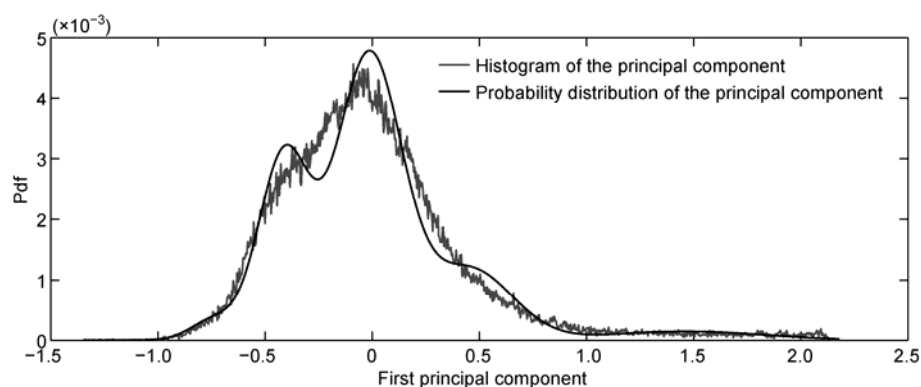
In this experiment, the first three principal components included 98% of the raw data (Figure 9). The peak of the first principal component was not significant, as shown by Figure 9(a). The class number (5) and initial centers (−0.8, −0.5, 0, 0.5, 1.5) were estimated based on the density function of the first principal component, as shown in Figure 9(a). Figures 10–12 further illustrate that the presented approach is reasonable compared to the class number and initial centers. Compared to Figure 5, the consistency of the histogram and the probability distribution of the first principal component are not very distinct.

As in the previous example, a comparison was made with the results (residence, farm1, farm 2, water, bare land) given by the *k*-means and modified algorithms. Figure 13(b)–(d) shows the classified images of the *k*-means algorithm, the modified EM method and the proposed method, respectively. The EM algorithm was superior to the *k*-means method, especially in the area mixed with multiple classes, such as

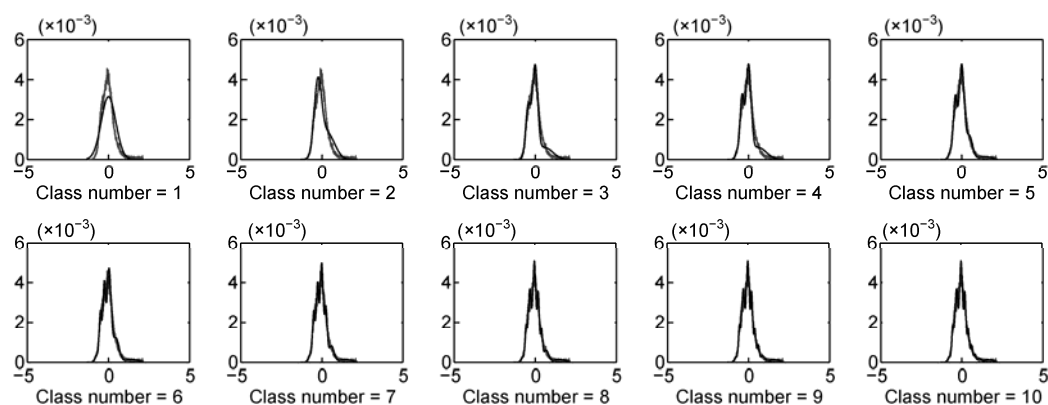




**Figure 9** Density function of the Log-Principal Component Transformation. Part (a) is the 1st Principal Component (The red dotted line corresponds to the center of each class). Parts (b) to (d), respectively, are the second to fourth principal components. The principal component is plotted along the  $x$ -axis, and the probability distribution is plotted along the  $y$ -axis.



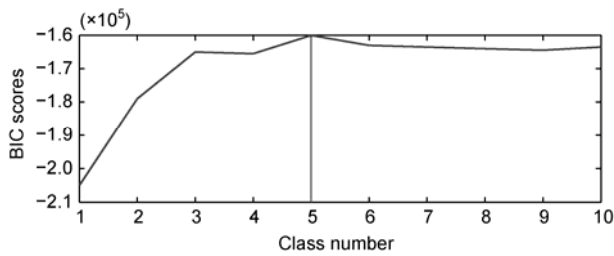
**Figure 10** Histogram comparing the probability distribution of the first principal component. The probability distribution was obtained by the proposed technique. The histogram was normalized to permit a direct comparison.



**Figure 11** Histogram and probability distribution of the first principal component. The probability distribution as a function of the class number  $k = 1, \dots, 10$ . The histogram was normalized to permit a direct comparison. The principal component is plotted along the  $x$ -axis, and the probability density function is plotted along the  $y$ -axis.

residential areas. The result of the modified EM algorithm was the best from 60 trials, similar to the proposed method. The modified EM method used the results of the  $k$ -means algorithm as initial centers. It is important to note that the  $k$ -means algorithm defines its initial parameters randomly,

so it is difficult to obtain optimal results from the modified EM method. Six hundred ground truth points randomly sampled on the reference image (Figure 13(a)) were verified by investigation. It was found that the overall accuracy and Kappa coefficient (Tables 5–7) obtained from the proposed



**Figure 12** The plot of BIC scores as a function of the class number for the first principal component.

approach are higher than those obtained from the other methods.

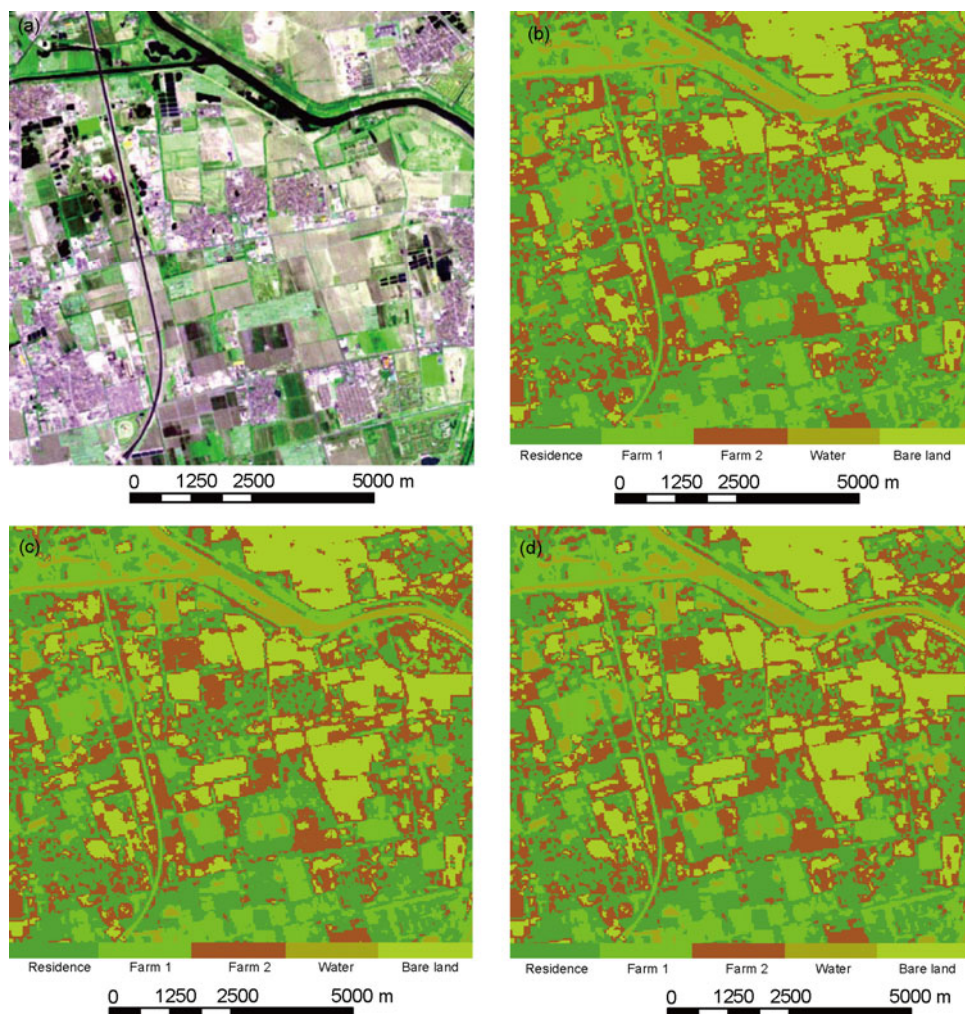
Compared with the previous examples, the quality and quantity of the classification results are lower. Some of the bands of the ASTER image are highly correlated. However, the spatial resolution of ASTER image is low. The first principal component from 9 bands of the ASTER image cannot fully reflect the characteristics of the classes. The resulting initializations are not very accurate, which affects

the final classification results. In conclusion, the proposed method is suitable for data sets with significant class features.

The modified EM algorithm uses the *k*-means results as an input. As is already known, the *k*-means algorithm is sensitive to the initial centers. Selecting the initial centers randomly can produce an unstable classification. Therefore, the best result must be chosen from several trials, which is a time-consuming process. Moreover, the modified EM algorithm cannot determine the class number. The proposed method not only gives reasonable class number and initial centers but also addresses the singularity problem by applying the principal component transformation. Therefore, we conclude that the stability of the proposed method is superior to that of the modified EM algorithm.

### 3 Conclusion

In this paper, we proposed a novel approach to improve the EM clustering algorithm. The presented approach is based



**Figure 13** (a) Pseudo color composite of band 2, 3N, and 1; (b) the *k*-means result; (c) the modified EM result; (d) the proposed algorithm result.

**Table 5** Classification accuracy of the proposed algorithm<sup>a)</sup>

	Water	Residence	Bare land	Farm 1	Farm 2	Sum
Water	43	0	0	0	0	43
Residence	0	74	14	8	21	117
Bare land	1	11	104	22	15	153
Farm 1	0	13	21	92	11	137
Farm 2	2	13	17	15	103	150
Sum	46	111	156	137	150	600

a) Overall classification accuracy: 69.33%; Kappa coefficient: 60.56%.

**Table 6** Classification accuracy of the modified EM algorithm<sup>a)</sup>

	Water	Residence	Bare land	Farm 1	Farm 2	Sum
Water	43	0	0	0	0	43
Residence	0	74	13	9	21	117
Bare land	1	11	101	24	16	153
Farm 1	0	9	24	88	16	137
Farm 2	2	14	19	18	97	150
Sum	46	108	157	139	150	600

a) Overall classification accuracy: 67.17%; Kappa coefficient: 57.76%.

**Table 7** Classification accuracy of the *k*-means algorithm<sup>a)</sup>

	Water	Residence	Bare land	Farm 1	Farm 2	Sum
Water	43	0	0	0	0	43
Residence	0	74	11	7	25	117
Bare land	1	11	99	24	18	153
Farm 1	0	9	27	87	14	137
Farm 2	2	15	23	19	91	150
Sum	46	109	160	137	148	600

a) Overall classification accuracy: 65.67%; Kappa coefficient: 55.83%.

on five theoretically well-founded methods for data and image analysis: the logarithmic transformation, the principal component transformation, the *k*-means algorithm, the EM algorithm and the Bayes classifier. A logarithmic transformation is used to obtain a set of symmetry data. Applying the principal component transformation to such data can result in a set of uncorrelated data. The initial parameters of the EM algorithm are determined by the *k*-means algorithm, whose centers are obtained from the density function of the first principal component. Such parameters are iteratively improved by applying the EM algorithm to obtain a more accurate set of parameters of the Gaussian mixture model. Finally, the resulting parameters are applied to a Bayes classifier to generate a classification map.

The proposed approach presents some important advantages over the general EM algorithm used in remote sensing applications: (1) The initial centers of the *k*-means algorithm are determined by analyzing the density function of the first principal component. In this way, the proposed approach avoids selecting random initial values. (2) The proposed approach not only solves the singularity but also

reduces the noise by using multiple principal components based on the criterion of the principal component. (3) In addition, log-transforming the multi-dimensional data enhances the type.

The experimental results reported in this paper highlight the effectiveness of the proposed approach. In particular, in all of the experiments that were conducted, the presented technique provided more accurate results than the *k*-means algorithm and the modified EM method. In addition, the experiments highlight the stability of the presented approach versus of the class number and initial centers. One of the main results drawn from the experiments is that the proposed method is suitable for data sets with significant category features.

In terms of computation time, the proposed method using multiple principal components is faster than the EM algorithm using a raw image because the EM algorithm takes a long time to process a large number of dimensions.

Finally, it is important to note that, although the proposed method was used for a remote sensing classification, it also exhibits properties that can be generalized to other applications. Hence, it can be used in any classification application requiring a technique based on the EM algorithm.

*This work was supported by the National High-tech R&D Program of China (2007AA12Z226 and SS2012AA120804), the National Natural Science Foundation of China (40674015 and 41074009), the Doctoral Fund of Ministry of Education of China (20100022110008), the Fundamental Research Funds for the Central Universities (2-9-2011-227), the Open Research Fund of Key Laboratory of Digital Earth Science, Center for Earth Observation and Digital Earth, Chinese Academy of Sciences (2010LDE002). The authors thank anonymous reviewers for giving valuable suggestion on the manuscript.*

- 1 Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. *J Royal Stat Soc*, 1977, 39: 1–38
- 2 Bruzzone L, Prieto D F. Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images. *IEEE T Geosci Remote S*, 2001, 39: 456–460
- 3 Figueiredo M A T, Jain A K. Unsupervised learning of finite mixture models. *IEEE T Pattern Ana*, 2002, 24: 381–396
- 4 Luo J C, Wang Q M, Ma J H, et al. The EM-based maximum likelihood classifier for remotely sensed Data (in Chinese). *Acta Geod E*, 2002, 31: 234–239
- 5 Chakravarty S, Qian D, Hsuan R. Adaptive gaussian mixture estimation and its application to unsupervised classification of remotely sensed Images. *Geoscience and Remote Sensing Symposium*, 2003. IGARSS'03. Proceedings. France: IEEE International, 2003, 3: 1796–1798
- 6 Kersten P R, Jong S L, Ainsworth T L. Unsupervised classification of polarimetric synthetic aperture radar images using fuzzy clustering and EM clustering. *IEEE T Geosci Remote S*, 2005, 43: 519–527
- 7 Korting T S, Dutra L V. Assessment of a modified version of the EM algorithm for remote sensing data classification. *Lect Notes Comput Sci*, 2010, 6419: 476–483
- 8 Yang H L, Peng J H, Li S H, et al. Log-principal component transformation based EM algorithm for remote sensing classification (in Chinese). *Acta Geod E*, 2010, 39: 378–382
- 9 Marzouki A, Delignon Y, Pieczynski W. Unsupervised statistical segmentation of multispectral SAR image using generalized mixture

- estimation. France: IEEE, 1996, 1: 706–708
- 10 Osamu U. A perceptive oriented approach to image Fusion. InTech I, 2011, 105–122
  - 11 Bruzzone L, Prieto D F. An adaptive semiparametric and context-based approach to unsupervised change detection in multitemporal remote sensing image. *IEEE Trans Image Process*, 2002, 11: 452–466
  - 12 Moser G, Serpico S B. Unsupervised change detection from multi-channel SAR data by markovian data fusion. *IEEE Trans Image Process*, 2009, 47: 2114–2128
  - 13 Fernández P, Marconcini M. A novel prtially spervised aproach to trgeted cange dtection. *IEEE T Geosci Remote Sens*, 2011, 99: 1–23
  - 14 Shi H, Cui H X, Yang H, et al. The exception maximization algorithm and its application in quantitative remote sensing inversion, Geoscience and Remote Sensing Symposium, 2004. IGARSS'04. Proceedings China: IEEE, 2004, 1: 641–644
  - 15 Reynolds D A, Rose R C. Robust text-independent speaker identification using gaussian mixture speaker models. *speech and audio processing. IEEE Trans*, 1997, 3: 72–83
  - 16 Bergasa L M, Mazo M, Gardel A, et al. Unsupervised and adaptive gaussian skin-color model. *Image and Vision Computing*. Amsterdam: Elsevier, 2000, 18: 987–1003
  - 17 Permuter H H, Francos J, Jarmyn I H. Gaussian mixture models of texture and colour for image database retrieval, Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03). German: IEEE, 2003, 3: 569–572
  - 18 Haim P, Joseph F, Ian J. A study of gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recogn*, 2006, 39: 695–706
  - 19 Fahrmeir L, Tutz G. *Multivariate Statistical Modeling Dased on Generalized Linear Models*. New York: Springer-Verlag, 1996
  - 20 McLachlan G J. *EM Algorithm and Its Extension*. New York: Wiley & Sons, 1997
  - 21 Jack T H. Population estimation models based on individual TM pixels. *Photogramm Eng Remote Sens*, 2002, 68: 1181–1192
  - 22 Chang C I, Du Q, Sun T L, et al. A joint band prioritization and band decorrelation approach to band selection for hyperspectral image classification. *IEEE Trans Image Process*, 1999, 37: 2631–2641
  - 23 Li Q L, Xue Y Q, Xiao D H, et al. Study on microscope hyperspectral medical imaging method for biomedical quantitative analysis. *Chin Sci Bull*, 2008, 53: 1431–1434
  - 24 Hye Y P, Tomoko O. Singularity and slow convergence of the EM algorithm for gaussian mixtures. *Neural Process Lett*, 2009, 29: 45–59
  - 25 Fwu J K, Djuric P M. EM algorithm for image segmentation initialized by a tree structure scheme. *IEEE Trans Image Process*, 1997, 6: 349–352
  - 26 Fraley C, Raftery A E. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *Computer J*, 1998, 41: 578–588
  - 27 Hartigan J A, Wong M A. A K-means clustering algorithm. *Appl Stat*, 1979, 8: 100–108
  - 28 Bazi Y, Bruzzone L, Melgani F. Image thresholding based on the EM algorithm and the generalized Gaussian distribution. *Pattern Recogn*, 2007, 40: 619–634
  - 29 Tyagi M, Bovolo F, Mehra A K, et al. A context-sensitive clustering technique based on graph-cut initialization and expectation-maximization algorithm. *IEEE T Geosci Remote Sens*, 2008, 5: 21–25
  - 30 Wen X B, Zhang H, Zhang J G, et al. Multiscale modeling for classification of SAR imagery using hybrid EM algorithm and genetic algorithm. *Prog Nat Sci*, 2009, 19: 1033–1036
  - 31 Ari C, Aksoy S. Maximum likelihood estimation of gaussian mixture models using particle swarm optimization. *Pattern Recognition (ICPR)*, 20th International Conference on Digital Object Identifier. IEEE, 2010, 746–749
  - 32 McLachlan G J, Peel D. *Finite Mixture Models*. NewYork: Wiley, 2000
  - 33 Hu Z Q, Xu C W. A new statistical method for mapping QTLs underlying endosperm traits. *Chin Sci Bull*, 2005, 50: 1470–1476
  - 34 Peng J H. Jointly robust estimation of unknown parameters and variance components based on expectation maximization algorithm. *J Surv Eng-Asce*, 2009, 135: 1–9
  - 35 Peng J H, Shi Y, Li S H, et al. MINQUE of variance-covariance components in linear gauss-markov models. *J Surv Eng*, 137: 129–139
  - 36 Bowman A W, Azzalin A. *Applied Smoothing Techniques for Data Analysis*. Oxford: Oxford University Press, 1997
  - 37 Fung T, LeDrew E. Application of principal component analysis to change detection. *Photogramm Eng Remote Sens*, 1987, 53: 1649–1658
  - 38 Xu B, Gong P. Land use/cover classification with multispectral and hyperspectral EO-1 data. *Photogramm Eng Remote Sens*, 2007, 73: 955–965
  - 39 Langley P, Iba W, Thomas K. An analysis of Bayesian classifiers. In: *Proceedings of the Tenth National Conference of Artificial Intelligence*. AAAI Press, 1992. 223–228
  - 40 Congalton R G, Mead R A. A quantitative method to test for consistency and correctness in photo-interpretation. *Photogramm Eng Remote Sens*, 1983, 49: 69–74
  - 41 Ma Z L, Redmond R L. Tau coefficients for accuracy assessment of classification of remote sensing data. *Photogram Eng Remote Sens*, 1995, 61: 435–439
  - 42 Yang H L. The Key Technology of SAR interferometry for detecting ground deformation studies. Dissertation for the Doctoral Degree. Beijing: China University of Geosciences (Beijing), 2012. 47–48

**Open Access** This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.